

A CENTRAL LIMIT THEOREM FOR REPEATING PATTERNS

ABSTRACT. This note gives a central limit theorem for the length of the longest subsequence of a random permutation which follows some repeating pattern. This includes the case of any fixed pattern of ups and downs which has at least one of each, such as the alternating case considered in [2] and [3]. In every case considered the convergence in the limit of long permutations is to normal with mean and variance linear in the length of the permutations.

1. SETUP

An r -**pattern** of **length** k is a map w from $\mathbb{Z}/k\mathbb{Z}$ to $2^{S_r} - \{\emptyset\}$ where S_r is a symmetric group. The case of up and down sequences is that of 2-patterns with every $w([a])$ either $\{\text{up}\} = \{(12)\}$ or $\{\text{down}\} = \{(21)\}$. A sequence $\sigma \in \mathbb{R}^n$ **follows** an r -pattern w if for every k the values $\sigma(k+1), \dots, \sigma(k+r)$ are distinct and the associated permutation ρ is in $w([k])$ where ρ is defined by having $\rho(i) < \rho(j)$ if $\sigma(k+i) < \sigma(k+j)$. Consider the uniform probability measure on each symmetric group S_n and regard S_n as a subset of $[1, n]^n \subseteq \mathbb{R}^n$. The main random variables of interest are L_n^w mapping S_n to \mathbb{Z}_+ with

$$L_n^w(\sigma) = \max\{r | \exists \tau : [1, r] \rightarrow [1, n] \text{ with every } \tau(i+1) > \tau(i) \text{ and } \sigma \circ \tau \text{ following } w\}$$

the length of the longest subsequence following w . A sequence $\{f_n\}$ of \mathbb{R} valued random variables **satisfies a clt** if there are $\mu \in \mathbb{R}$ and $0 < \sigma \in \mathbb{R}$ With

$$\lim_{n \rightarrow \infty} \text{Prob}\left(f_n - \mu n < t\sigma\sqrt{n}\right) = \Phi(t)$$

for every $t \in \mathbb{R}$. Here $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{u^2}{2}} du$ is the cumulative distribution function for the standard normal distribution.

The fact that every nonconstant up and down pattern w has $\{L_n^w\}$ following a clt will follow as a corollary of the theorem below. The proof does not give good bounds on the means or variances involved. This is in contrast to the special case of alternating subsequences, where $\mu = \frac{2}{3}$ and $\sigma^2 = \frac{2}{9}$ are straightforward to compute.

Write $\sigma\tau = (\sigma(1), \dots, \sigma(a), \tau(1), \dots, \tau(b))$ for the product and $\sigma^m = \sigma \dots \sigma$ for the m th power if $\sigma \in \mathbb{R}^a$ and $\tau \in \mathbb{R}^b$ are sequences of lengths a and b . The proof will also require the patterns to be combinatorial, though it is not clear whether this is the best condition possible. In particular the constant up and down patterns do not follow a clt so some condition is needed.

Call a pattern w **combinatorial** if every pair of sequences of lengths a and b following w have a subsequence of their product of length at least $a+b-k$ which also follows w , where k is the length of w .

2. THEOREM

Theorem. If w is a combinatorial pattern then $\{L_n^w\}$ satisfy a clt.

The following is easy:

Corollary. Every nonconstant 2-pattern is combinatorial and hence follows a clt.

Note that there are also other nontrivial examples. For instance the constant 3-pattern of length 3 with every $w([i]) = \{(123), (231), (312)\}$ is combinatorial.

3. PROOF

The idea is to switch from uniformly chosen permutations to sequences of independently selected points in an interval. In this setting one can choose a positive probability event depending on only a short subsequence of points so that the longest subsequence following the given pattern is found by combining a longest one before the event and a longest one after it. The problem is thus reduced to a sum of independent events with good enough control on the number of these events.

Consider the probability space $\Omega = [0, 1]^{\mathbb{Z}_+}$ with the product Lebesgue measure. Write $\pi_A : \Omega \rightarrow [0, 1]^A$ for the projections and $R_{[a,b]}^w$ for the random variable with

$$R_{[a,b]}^w(\sigma) = \max\{|A| : A \subseteq [a, b] \text{ and } \pi_A(\sigma) \text{ follows } w\}$$

the length of the longest subsequence in the interval following w and $R_n^w = R_{[1,n]}^w$ for the family of random variables of interest. Note that the push forward of the uniform measure under L_n^w and that of the Lebesgue measure under R_n^w are the same measure on \mathbb{Z}_+ and hence one family satisfies a clt iff the other one does.

Fix a combinatorial r -pattern w of length k .

In the notation of [1], if w drifts up then taking a to be a long sequence following w with values above $\frac{1}{2}$ and b one with values below $\frac{1}{2}$ would show that w is not combinatorial so w is driftless. By Proposition 4.10 of [1] this gives a totally driftless loop which in turn gives a permutation $\tau \in S_k$ so that every power of τ follows w .

The decomposing event is the subcube B of $[0, 1]^{4k}$ with volume $\text{vol}(B) = (3k)^{-4k}$ given by

$$B = \left(\prod_{i \in [1, k]} \left(\frac{-2}{3k} + \frac{\tau(i)}{k} + \frac{1}{3k}[0, 1] \right) \right)^4.$$

Note that every element of B follows w .

Here are some random variables which decompose the sequences at the B events. Take $\{d_j\}$ for $j \geq 1$ to be the iid indicator random variables for B so that

$$d_j(\sigma) = \begin{cases} 1 & \text{if } \pi_{(4kj-4k, 4kj]}(\sigma) \in B \\ 0 & \text{otherwise} \end{cases}$$

with means $\mu_d = \text{vol}(B)$ and variances $\sigma_d^2 = \text{vol}(B)(1 - \text{vol}(B))$. For the partial sums write $D_A = \sum_{j \in A} d_j$ and $D_n = D_{[1, n]}$. Next build approximate inverse random variables to D_n using for $j \geq 0$ the iid random variables $\{p_j\}$ with

$$p_j(\sigma) = \min\{a_j \mid \text{there are } 0 \leq a_0 < \dots < a_j \text{ with every } d_{a_i} = 1\}$$

for the position of the $(j+1)$ st occurrence of an event in B and for $j \geq 1$ the iid variables $\{q_j\}$ with

$$q_j = p_j - p_{j-1}$$

for the number of steps between events in B . Here a sequence of $4kn$ elements in $[0, 1]$ has n positions and adjacent positions differ by a single step. Note that these are defined almost everywhere in Ω . Again write Q_A and Q_n for the associated partial sums. Thus $\mu_q = \mu_d^{-1} = (3k)^{4k}$, Q_{D_n} is the first position of an event in B after position n and $D_{p_n} = n + 1$.

Here are some random variables involving those above and lengths of subsequences following w . For $j \geq 1$ take the iid random variables $\{s_j\}$ to be

$$s_j = R_{(4kp_{j-1}+2k, 4kp_j+2k]}^w$$

for the length of the longest subsequences between the midpoints of successive events in B which follow w . Take $\{t_j\}$ to be

$$t_j = s_j - \mu_s \mu_d q_j$$

another collection of iid random variables. Note that since $4k \leq s_j \leq 4kq_j$ the means μ_s are finite and hence the t_j are also defined almost everywhere and have means $\mu_t = 0$. Since there are many different positive probability values for the t_j their variances σ_t^2 are nonzero and since $-\mu_s \mu_d q_j \leq t_j \leq s_j$ they are also finite.

Note that since w is combinatorial and every event in B follows w , every longest subsequence of σ following w includes every interval $(4kp_j + k, 4kp_j + 3k]$, which is the middle $2k$ indices of an event in B and similarly every longest subsequence of the projection $\pi_{(4kp_i+2k, 4kp_j+2k]}(\sigma)$ following w includes the end intervals $(4kp_i + 2k, 4kp_i + 3k]$ and $(4kp_j + k, 4kp_j + 2k]$. This implies

Lemma 1. $R_{(4kp_1+2k, 4kp_{D_n}+2k]}^w = S_{D_n}$.

This sum is further decomposed into (2)+(3)+(4) for easier analysis and the short ends (1) and (5) are added to get

Lemma 2. $R_{4kn}^w = (1) + (2) + (3) + (4) + (5)$ where

$$(1) = R_{4kp_0+2k}^w,$$

$$(2) = T_{\lfloor \mu_d n \rfloor},$$

$$(3) = \begin{cases} T_{(\lfloor \mu_d n \rfloor, D_n]} & \text{if } \lfloor \mu_d n \rfloor \leq D_n \\ -T_{(D_n, \lfloor \mu_d n \rfloor]} & \text{otherwise} \end{cases},$$

$$(4) = \mu_s \mu_d Q_{D_n}$$

and

$$(5) = -R_{(4kn, 4kp_{D_n}+2k]}^w.$$

For the desired large n limit the terms (1), (3) and (5) are too short to contribute, though for (3) this uses Kolmogorov's inequality. The mean comes from (4), which has vanishing variance and the variance comes from (2) with vanishing mean.

(1),(5): Since $|(1)| \leq 4kp_0$ which is independent of n there is

$$\lim_{n \rightarrow \infty} \text{Prob}\left(|(1)| > a\sqrt{4kn}\right) = 0$$

for every $a > 0$ and similarly for (5).

(3): Assume the first case in the definition of (3) holds. The second case is similar. The number of t_j terms in the sum is $|D_n - \lfloor \mu_d n \rfloor|$ and since D_n is a sum n of the d_j iid random variables the central limit theorem gives

$$\lim_{n \rightarrow \infty} \text{Prob}\left(|D_n - \lfloor \mu_d n \rfloor| > v\sigma_d\sqrt{n}\right) = \Phi(-v)$$

and Kolmogorov's inequality gives gives for every a and b that

$$\text{Prob}\left(\max_{j < a} |T_{(\lfloor \mu_d n \rfloor, \lfloor \mu_d n \rfloor + j]}| > b\right) \leq \frac{a\sigma_t^2}{b^2}.$$

Taking $a = n^{\frac{1}{2}+\epsilon}$ and $b = n^{\frac{1}{2}-\epsilon}$ for any $0 < \epsilon < \frac{1}{6}$ gives

$$\text{Prob}\left(|(3)| > \frac{\sqrt{n}}{n^\epsilon}\right) \leq \sigma_t^2 n^{3\epsilon-\frac{1}{2}} + \text{Prob}\left(|D_n - \lfloor \mu_d n \rfloor| > n^\epsilon \sqrt{n}\right)$$

and hence

$$\lim_{n \rightarrow \infty} \text{Prob}\left(|(3)| > \frac{\sqrt{n}}{n^\epsilon}\right) = 0 + \lim_{n \rightarrow \infty} \Phi\left(-\frac{n^\epsilon}{\sigma_d}\right) = 0.$$

(4): Since Q_{D_n} is the position of first occurrence after n of an event in B , $\text{Prob}\left(Q_{D_n} - n \geq a\right) = (1 - \mu_d)^{\lceil a \rceil}$. Thus for any $u > 0$

$$\lim_{n \rightarrow \infty} \text{Prob}\left(|(4) - \mu_s \mu_d n| \geq u \sqrt{n}\right) = \lim_{n \rightarrow \infty} \text{Prob}\left(Q_{D_n} - n \geq \frac{u \sqrt{n}}{\mu_s \mu_d}\right) = 0.$$

(2): Since the second term is a sum of $\lfloor \mu_d n \rfloor$ iid mean zero variables t_j , the central limit theorem gives for every u that

$$\lim_{n \rightarrow \infty} \text{Prob}\left((2) < u \sigma_t \sqrt{\mu_d n}\right) = \Phi(u).$$

Adding these gives for every u the desired

$$\lim_{n \rightarrow \infty} \text{Prob}\left(R_{4kn}^w - \mu_s \mu_d n < u \sigma_t \sqrt{\mu_d n}\right) = \Phi(u).$$

4. FURTHER DIRECTIONS

The mean length for any pattern is roughly controlled by the drift in that patterns with drift have a mean length of order \sqrt{n} by a comparison to the increasing case, while those without have mean linear in n . It is not clear whether the combinatorial property is needed for a central limit theorem.

REFERENCES

1. Aaron Abrams, Eric Babson, Henry Landau, Zeph Landau, James Pommersheim, *Distributions of Order Patterns of Interval Maps*, arXiv:1003.5561.
2. Richard Stanley, *Longest Alternating Subsequences of Permutations*, arXiv:0511419.
3. Harold Widom, *On the Limiting Distribution for the Length of the Longest Alternating Sequence in a Random Permutation*, arXiv:0511533.

Authors: Aaron Abrams, Eric Babson, Henry Landau, Zeph Landau and James Pommersheim.